DOI:10.20176/j.cnki.nxdz.20260703

机器学习在大气污染物浓度预测中的研究进展

周士杰[1], 王 刚[1*], 赵 娜[2]

(1.中国石油大学(华东) 化学化工学院,山东 青岛 266580; 2.山东大学 环境研究院,山东 青岛 250100)

摘 要:大气污染物浓度的精准预测是大气污染防治的重要前提。传统预测方法在计算成本、非线性拟合能力及泛化性能等方面存在固有局限,难以满足复杂大气环境下的高精度预测需求。近年来,机器学习技术的快速发展为该领域提供了新的解决方案。系统梳理用于大气污染物浓度预测的机器学习方法,并将其归纳为传统机器学习(如支持向量回归、随机森林、梯度提升树等)与深度学习方法两大基本范畴。对于深度学习方法,又依据其处理任务不同,将其划分为时间序列模型、空间尺度模型及时空融合模型;从模型机理、预测性能、适用场景及优缺点等维度,对各类模型进行比较分析。结果表明,传统机器学习方法在站点尺度短中期($t\leqslant 72$ h)预测中表现稳定,具有良好的适用性;深度学习方法中,时间序列模型可有效提升多时间尺度预测精度,空间尺度模型提高了对空间关联特征的捕捉能力,而时空融合模型则可实现时间依赖性与空间关联性的协同建模,展现出较强的综合预测性能。未来该领域模型发展将朝着多源数据融合、模型架构轻量化与集成化、预测结果可解释化的方向推进;结合迁移学习、联邦学习等技术解决区域监测数据稀缺问题;拓展、提高模型在跨区域与极端污染事件等场景下的适用范围、预测精度,从而为大气污染的精准防控提供有力技术支撑。

关键词:大气污染物预测;传统机器学习;深度学习

分类号:(中图)X513;TP181 **文献标志码**:A

大气污染是当前全球最受关注的环境问题之一[1-2],其来源既包括工业、交通等领域直接排放的一次污染物,也包括复杂气象条件下一次污染物生成的光化学二次污染物。这些污染物对生态环境、经济发展及人类健康构成持续威胁,尤其对人体呼吸系统、心血管系统等多系统功能造成长期危害,已成为全球公共卫生领域面临的重要挑战[3-4]。因此,精准预测大气污染物浓度,是有效遏制空气污染、保护公众健康、优化城市布局与解析气候效应的关键前提[5]。研究结果显示,提高细颗粒物($PM_{2.5}$)的预报精度,能够显著提高健康风险预警能力,帮助公众及时规避污染暴露风险[6]。传统大气污染物浓度预测方法主要分为基于物理化学机制的数值模型(如CMAQ[7]、WRF-Chem[8])与基于统计分析的统计模型(如ARIMA[9]、多元回归[10]):前者依赖详尽的排放清单与复杂的参数化方案,常面临计算成本高、实时性弱及对排放源不确定性敏感等问题;后者则通常假设变量间呈线性关系,难以有效刻画污染物与气象、地理等多因素间的非线性相互作用,两者均对数据质量有较高要求。

在此背景下,机器学习方法凭借其强大的非线性拟合能力、高效的自动特征提取机制及对高维数据的建模优势,为大气污染物浓度预测提供了新的技术路径,并逐渐成为该领域的重要研究方向[11]。机器学习方法中深度学习方法表现出较好的综合预测性能,这在大气污染物浓度预测中尤为重要[5, 12-13]。近年来,为提升预测性能与模型适应性,机器学习逐渐向多模型融合与混合架构方向发展,通过集成不同模型的优势,实现对污染物时空演变规律更精准、稳健的刻画,从而推动其向智能化、集成化及可解释化方向发展[14]。为了满足大气污染防控与环境管理的实际需求,污染物浓度的精准预

收稿日期:2026-01-24

基金项目:国家自然科学基金资助项目(42507143)

作者简介:周士杰(2002—),男,硕士研究生,主要从事机器学习在大气污染中的应用研究,(电子信箱)zhousj6074@163.com。

*通信联系人:王刚(1988—),男,副教授,博士,主要从事大气复合污染成因与来源研究,(电子信箱)wangg@upc.edu.cn。

引用格式:周士杰,王刚,赵娜.机器学习在大气污染物浓度预测中的研究进展[J].宁夏大学学报(自然科学版中英文),2026,47(4):302-310.

测已成为关键技术难题。传统预测方法存在计算量大、非线性表达能力不足、泛化能力有限等问题，难以完全满足实际防控工作对预测精度的要求。尽管机器学习已成为该领域的重要技术支撑，但传统机器学习与深度学习方法种类繁多，各种模型的适用场景、性能特点与优劣存在明显差异，且对现有研究缺乏系统梳理，导致使用者在实际应用中难以针对具体问题高效选择合适的方法[15-16]。为此，文中系统综述大气污染物浓度预测领域的主流机器学习方法，旨在明确各模型的适用条件、核心性能与优缺点，为同类区域的污染物浓度精准预测提供技术参考。

1　机器学习在大气污染物浓度预测中的应用

1.1　传统机器学习方法

在预测大气污染物浓度时，传统机器学习方法的核心原理是基于统计学理论，即从历史监测数据（包括污染物浓度、气象参数、排放源信息等）中挖掘输入特征与污染物浓度之间的非线性映射关系，构建拟合模型并完成预测。其中，支持向量回归（support vector regression，SVR）、随机森林（random forest，RF）及梯度提升树（gradient boosting decision tree，GBT）等模型被广泛应用，且取得了显著成效[17]。

1.1.1　SVR模型　SVR模型通过核函数将低维数据映射至高维特征空间，构造拟合数据的最优超平面。该方法通过正则化机制控制模型复杂度，有效避免了过拟合，并在处理小样本、高维度数据时，仍能较好地捕捉变量间的非线性关系。然而，在处理大规模数据时，SVR模型的计算效率较低，且其预测性能受污染物类型、时空尺度、数据质量及优化策略的综合影响。因此，近年来的研究多围绕优化SVR模型展开。例如，采用基于线性核[18]、径向基函数（radial basis function，RBF）核[19]等不同形式的SVR模型，或将SVR模型与其他算法结合，已成为实现污染物浓度与空气质量指数（air quality index，AQI）精准预测的常见方法[20]。部分研究重点改进SVR模型的参数设置，通过引入群智能或元启发式搜索算法自动调整关键参数，不仅降低了模型误差，也使预测结果在多次实验中更集中、波动更小且表现出较好的稳定性与鲁棒性[21]。还有研究以核SVR模型为主体，辅以数据预处理技术，在小时尺度、小样本等条件下仍能保持较强的回归拟合能力，并将AQI分级预测转化为分类任务，从而得到较高的准确率[22]。此外，有研究引入RF模型用于特征筛选，以消除冗余变量，并在输入端添加时间滞后项，以刻画污染物的动态延续性，从而在计算量增加不大的前提下，显著提升测试集的预测效果[23]。总体而言，参数自适应、特征筛选与时序信息引入，是优化SVR模型并将其用于空气质量预测最常见且有效的3种路径。主流研究结果的平均绝对百分比误差（MAPE）为0.3%～35.0%，部分场景决定系数（R^2）可达0.83～0.99（表1，表1中e_{map}、e_{rms}、e_{ma}、e_{mr}分别表示平均绝对百分比误差、均方根误差、平均决定误差、平均相对误差）

1.1.2　RF模型　RF模型基于集成学习思想，通过Bootstrap重采样与随机特征选择构建多棵决策树，并以投票或平均策略方式输出结果。该模型具备较强的抗过拟合能力，对数据缺失和异常值具有良好的鲁棒性，能高效处理非线性关系与高维数据。然而，在大规模长时序数据场景下，RF模型的计算资源消耗较大，且模型的可解释性相对较弱。因此，RF模型更适用于站点级与区域级、短期至中期（$t\leqslant$72 h）污染物预测。目前，主流研究常将RF回归与可解释性框架或预测软件相结合。部分研究针对历史监测数据缺乏的区域，通过融合遥感反演产品、气象观测及再分析资料，构建基于RF的污染物浓度重建模型。这些模型在月、季、年等多个时间尺度上均表现出稳定的预测性能，并成功生成高空间分辨率的长期污染数据集[24]。还有研究则根据高频监测数据，构建分辨率为秒级或分钟级的预测模型，该类模型在实现回归拟合精度较高的同时，亦能解析气象要素与前体物对污染物生成的协同影响机制[25]。此外，RF模型结合SHAP（SHapley additive exPlanations）等可解释性分析方法，不仅实现对污染物浓度的精准预测，还能有效识别关键影响因子及其交互作用，从而在提升模型预测性能的同时增强其物理可解释性，为污染成因分析与控制决策提供量化依据[26]。目前主流研究模型的R^2普遍在0.75以上，其中，对多数城市站点的$PM_{2.5}$、O_3、NO_2等污染物进行预测，R^2=0.75～0.90，e_{map}=8%～18%，且在部分污染物浓度较高时段，e_{map}降至5%以下（表1）。

1.1.3　GBT系列模型　GBT系列模型基于加法模型与前向分步算法，通过迭代训练弱决策树拟合来逐步降低预测误差。其中，极致优化的梯度提升树XGBoost（extreme gradient boosting）通过引入正则化项增强抗过拟合能力，而轻量级梯度提升树LightGBM则借助直方图算法与梯度单边采样法提

表1 各模型预测性能对比

Tab. 1 Comparison of prediction performance of each model

模型类别	模型	适用场景	样本量/条	关键性能指标	优势	局限性	参考文献
传统机器学习	SVR	站点级小样本、高维数据预测	$10^3 \sim 10^4$	e_{map}=0.3%～35.0%，R^2=0.83～0.99	泛化能力强，不易过拟合	计算效率低，对异常值敏感	[21-23]
	RF回归	站点级、区域级短期至中期($t \leqslant 72$ h)预测	$10^4 \sim 10^6$	$PM_{2.5}$、O_3、NO_2预测中$R^2 \geqslant 0.75$，且R^2集中在0.75～0.90；e_{map}=8%～18%，在高浓度时段，$e_{map}<5\%$	抗过拟合能力强，对缺失值、异常值鲁棒性强	极端污染预测易出现偏差，计算资源消耗大	[24-26]
	GBT系列	区域级污染物优化模拟	$10^4 \sim 10^6$	$R^2 \geqslant 0.75$，e_{map}=8%～18%	抗过拟合能力强，精度高	计算复杂度随着样本量上升，可解释性弱	[27-30]
时间序列	LSTM	短中期(t=6～336 h)污染物预测	$\geqslant 10^4$	短期预测，R^2=0.70～0.97，48 h仍保持$R^2>0.60$，e_{ma}=5～15 μg/m³，e_{rms}=8～25 μg/m³；高污染事件命中率为56%～92%	长时序依赖捕捉能力强，结构成熟，调参难度低	长序列训练耗时长，空间关联捕捉能力弱	[13,31-33]
	Transformer	中长期($t>72$ h)多周期预测	$\geqslant 10^5$	R^2=0.81～0.94，e_{ma}=1.39～5.30 μg/m³，e_{rms}=2.08～8.25 μg/m³	跨尺度特征融合能力强，并行效率高	参数规模大，依赖海量数据与大量计算资源	[34-35]
空间尺度	CNN	区域级高分辨率(10 m)网格浓度预测	$\geqslant 10^4$	$PM_{2.5}$预测，R^2=0.85～0.95；O_3预测，R^2=0.75～0.90	空间特征提取能力强，计算效率优于数值模拟	非欧氏空间数据适配能力低，时序依赖捕捉能力弱	[36-39]
	GNN	区域级多站点协同预测	$\geqslant 10^4$	R^2=0.75～0.95，e_{rms}=2.08～6.14 μg/m³，e_{ma}=1.39～5.00 μg/m³	适配非规则空间，精准捕捉跨区域传输效应	计算复杂度与站点成正比	[40-43]
时空融合	CNN-LSTM	市级多站点小时级、日级预测	$\geqslant 10^4$	O_3预测，$R^2 \geqslant 0.85$，$PM_{2.5}$预测，R^2=0.66～0.90，e_{rms}=9.66～24.65 μg/m³	同步捕捉时空规律，复杂场景鲁棒性强	非规则站点数据适配能力低，长距离传输捕捉能力弱	[31,44-47]
	GNN-LSTM	省域、跨区域中期(t=24～72 h)预测	$\geqslant 10^4$	$PM_{2.5}$、O_3的72 h预测，e_{mr}=17.7%～22.8%，R^2=0.64～0.90，e_{ma}=4～27 μg/m³，e_{rms}=8～27 μg/m³	复杂地形适配优，长时预测稳定，极端污染事件召回率高	大规模站点训练效率低，跨区域泛化时需重新优化结构	[41,48-50]

升训练效率[27]。该类模型的核心优势在于兼具高预测精度与高计算效率，既能精准捕捉污染物与多源因子之间的复杂非线性关系，又支持特征重要性评估。然而，该类模型对超参数调优比较敏感，通常需要通过大量实验确定最优配置，且训练过程中易受异常值干扰。梯度提升框架下的多种算法，已被广泛用于多尺度、高分辨率污染物浓度模拟与数据集重构。相关研究通过引入空间特征或结合时序波动性建模，显著提升了该类模型的预测精度与泛化能力，可适用于多特征耦合的复杂污染场景，以及站点级、区域级短期至中期($t \leqslant 72$ h)预测。例如，Zhong等[28]基于全国气象站点数据，构建融合空间特征的LightGBM模型，实现了从小时到年尺度的$PM_{2.5}$浓度预测，并在独立区域验证中表现出较好的空间泛化性能。另有研究采用LightGBM模型对中国大陆$PM_{2.5}$长期浓度进行重构，所得数据集的相关系数与均方根误差等指标，均显著优于原始再分析产品及其他树模型的预测结果。此外，利用XGBoost对局部地区的$PM_{2.5}$、O_3浓度进行优化模拟，可生成接近观测数据的高时空分辨率数据集[29]。还有研究进一步将树模型与神经网络、波动性模型相结合，构建混合预测框架。实验结果显

示，该模型在保证预测精度的同时，显著增强了对噪声的鲁棒性及对未参与训练站点的泛化能力[30]。主流预测研究模型的 $R^2 \geqslant 0.75$，$e_{map}=8\%\sim18\%$（表 1）；在站点级小时至日尺度预测中，GBT 系列模型仍是首选模型。

1.2　深度学习方法及融合技术

在大气污染物浓度预测中，深度学习方法的核心原理在于，通过多层神经网络自动提取多源数据（包括监测数据、气象资料、遥感影像等）的高维时空特征，构建污染物浓度与各类影响因子之间复杂的非线性及时空耦合关系。该方法无需人工预设特征映射规则，适用于大规模、多维度的污染预测场景[11]。根据建模的侧重点，深度学习模型可分为时间序列模型、空间尺度模型以及时空融合模型。在实际应用中，可根据具体预测需求及场景特点选择合适的模型。

1.2.1　*时间序列模型*　时间序列模型的核心原理在于，通过循环机制或注意力机制挖掘污染物浓度序列数据中的动态关联特征，有效捕捉昼夜节律、季节周期及滞后效应等时间依赖特性。常用模型主要包括长短期记忆神经网络（long short term memory，LSTM）、门控循环单元（gated recurrent unit，GRU）以及基于 Transformer 的架构[51-52]。其中，LSTM 模型通过门控机制，有效缓解传统循环神经网络中的梯度消失问题，能够较好地捕捉污染物浓度的长时序依赖特征。该模型结构成熟、调参难度相对较小，在短中期尺度（$t=6\sim72$、$72\sim336$ h）的污染物浓度预测中表现出显著适用性[31]。相关研究大多通过模型衍生、结构优化及多源数据融合等方法提升模型性能[13]。例如，有研究基于双向 LSTM 模型并结合注意力机制，实现延伸期污染过程持续时段的预测；亦有研究通过引入多源辅助数据与多分支输入架构，构建融合气象因子及卫星校正火情干扰信息的 LSTM 模型，该模型在天级尺度预测中取得优于传统机器学习方法的较高精度。然而，LSTM 模型在处理超长序列时仍面临训练效率下降的挑战，且对空间关联的建模能力有限。因此，当前研究多倾向于将 LSTM 模型与卷积神经网络、图神经网络等具备空间特征提取能力的模块相结合，构建更具表达力的时空联合预测模型[32-33]。

GRU 模型作为 LSTM 模型的轻量化变体，通过更新门与重置门的协同作用，高效捕捉大气污染物浓度的时序依赖特征，进而使参数规模更加精简。因此，GRU 模型具备训练效率高、计算成本低的优势，更适用于超短期（$t\leqslant6$ h）实时污染物浓度预测，目前主要被用于 $PM_{2.5}$ 浓度预测。Becerra-Ric 等[53]通过对比实验，验证了 GRU 模型在长时序预测中的稳定性，发现在 120 h 的 PM_{10} 浓度预测中，GRU 模型表现稳健，最低 $e_{rms}=3.1\ \mu g/m^3$，但该模型的可解释性相对有限。为同时提升模型的时空预测能力与可解释性，Guyu 等[54]提出融合图卷积与自注意力机制的 MGCGRU-SAN 模型，该模型在京津冀地区 $PM_{2.5}$ 浓度短期预测中表现优异（1～3 h 预测的 $e_{ma}\leqslant8\ \mu g/m^3$），且在 48 h 预测中保持误差稳定（$e_{rms}<0.3$），有效弥补了传统方法在空间关联与长期依赖建模方面的不足。

Transformer 框架依托自注意力机制，能够同步捕捉大气污染物浓度、气象因子等多源数据的全局关联特征与长短期时序依赖关系，并支持并行训练，且其计算效率优于传统递归模型[34]。因此，Transformer 框架更适用中期（$t=72\sim336$ h）及长期（$t>14$ d）的多周期预测任务。为应对预测中存在的复杂时空依赖与噪声干扰问题，混合架构的创新成为关键。例如，Wu 等[35]通过数据标准化与多源数据融合策略，显著提升了单一 Transformer 模型在 $PM_{2.5}$ 浓度预测中的性能（R^2 提升至 0.94）。然而，此类模型通常参数量大，且依赖高质量的大规模数据与较强的算力支撑。

1.2.2　*空间尺度模型*　空间尺度模型的核心在于挖掘污染物浓度的空间关联特征（如区域传输、站点间扩散效应），并通过卷积操作或图结构建模的方式，进而适配非欧氏空间的特征分布[55]。该类模型主要包括卷积神经网络（convolutional neural network，CNN）[36-37]与图神经网络（graph neural network，GNN）[40-41]，两者凭借差异化的空间建模能力，分别适用于不同场景的预测需求。

CNN 模型适用于区域级网格化浓度预测，尤其在结合卫星遥感气溶胶光学厚度（aerosol optical depth，AOD）数据进行污染物空间分布反演等场景中表现突出。该模型的优势在于强大的空间特征自动提取能力，即能够高效捕捉大气污染物与地形、建筑、排放源等影响因素的非线性关联关系。例如，Wang 等[38]融合并行化大涡模拟模型（parallelized large-eddy simulation model，PALM），模拟数据与城市形态数据，构建用于预测南京城区 10 m 级 CO 浓度的 CNN 模型。研究结果显示，在多种情景下该模型的 R^2 达 0.86，其计算效率较原物理模型提升 3 个数量级，并能通过权重分析量化建筑与排放源的影响。Chen 等[39]提出 CNN-RF 集成框架，即

先通过CNN提取深层特征,再由RF完成回归分析。在高雄地区$PM_{2.5}$浓度预测中,该模型的$R^2=0.88$,其稳定性在不同浓度阈值下均优于单一模型。然而,CNN在处理非规则分布站点数据和长时序依赖关系方面存在局限,其"黑箱"特性也限制了对单一影响因子的定量解析能力。

GNN模型则更适用于区域级多站点协同预测,尤其在地形复杂、站点分布不均区域的污染物传输(如山地与城市间$PM_{2.5}$的跨区域扩散)模拟中具备独特优势。相关研究主要围绕架构融合与建模范式突破两个方向展开。例如,Liao等[42]创新性采用空气质量轨迹模型HYSPLIT(hybrid single-particle lagrangian integrated trajectory model)进行轨迹分析,以此动态构建图节点间的边连接关系,从而模拟污染物的实际传输路径;同时,结合注意力门控循环单元实现时序特征提取。研究结果显示,在1~48 h预测中,该模型的e_{ma}与e_{rms}较先进GNN基线,分别平均降低约14.0%、13.0%。Manda等[43]提出一种统一谱域与空间域的建模范式,即统一谱-空间图神经网络(unified spectral-spatial graph neural network, USS-GNN),并利用两者间光化学转化的内在耦合关系,构建O_3与NO_2的双变量同时预测框架。相比单变量模型,在O_3、NO_2浓度预测中,该模型的R^2分别提升8.5%、6.6%。

1.2.3 时空融合模型 时空融合模型旨在同步捕捉污染物浓度的时间依赖性与空间关联性,通过多模块协同建模实现时空耦合特征的深度挖掘[56-57]。其代表性模型主要包括CNN-LSTM模型[31,44]与GNN-LSTM模型[41,44,48],两者凭借差异化时空特征提取机制,在不同尺度与场景预测中展现出各自优势。

CNN-LSTM模型采用"空间特征提取-时序动态挖掘"的协同架构,已成为城市污染物时空预测的主流范式。相关研究围绕架构优化与多源数据深度耦合这两个方向持续拓展其性能。在架构优化方面,引入注意力机制可有效提升模型对关键时空特征的聚焦能力,并在O_3浓度预测中实现$R^2=0.97$[45]。在数据耦合方面,融合区域气象场的时空演变序列(如GFS/ERA5[46]),使模型能够物理性刻画台风等大型天气过程对污染物浓度的影响,将区域O_3浓度预测的R^2提升至0.93[47]。该模型在同步刻画局地污染梯度与日变化规律方面表现优异,但其预测结果难以分解为单一影响因子的定量贡献。同时,在大规模监测网络下,该模型的训练效率偏低。

针对空气质量监测站点非规则分布及污染物跨区域传输问题,GNN-LSTM模型基于图结构实现对空间依赖关系的动态建模,为污染物的精准模拟提供了较优的解决方案。相关研究聚焦于图结构精细化构建与表达,从基于地理距离的静态图发展到融合气象场(如风场)的动态有向图,通过引入物理约束提升污染物传输模拟的准确性。此类模型在模拟北京地区$PM_{2.5}$、O_3的跨站点传输时,可将72 h预测的e_{mr}控制在22.8%以内[49]。进一步引入空间聚类与图注意力机制,可使模型自适应识别具有相似污染模式的城市子区域,从而在复杂城市群中实现短期多污染物的协同预测($R^2>0.85$)[50]。然而,计算效率受站点规模影响较大,图结构高度依赖区域地理气象特征,使该类模型的跨区域泛化能力仍面临较大挑战。

此外,亦有研究提出GNN-Transformer等新型混合架构模型。例如,Ye等[58]通过引入瞬时相位同步估计器,自适应筛选关联站点数据并降低噪声,进而实现了深度时空特征的提取与预测。研究结果显示,该模型在短中期($t=3\sim24$ h)污染物浓度预测中表现优异($R^2=0.97$),尤其对高浓度$PM_{2.5}$的预警具有较高可靠性。

综上所述,以SVR、RF、GBT为代表的传统机器学习模型,凭借在小样本场景下的良好泛化能力、对非线性关系的有效捕捉和较强的可解释性,在站点级短中期预测中展现出优异性能。以长短期记忆网络、卷积神经网络、图神经网络及其融合模型(如CNN-LSTM、GNN-LSTM)为核心的深度学习方法,则可自动提取高维时空特征,在刻画污染物的复杂时空耦合机制及实现区域级高分辨率预测方面,取得突破性进展。各类模型在适用场景、核心性能与计算效率上呈现互补特征,对其持续优化与融合创新,有助于推动该领域技术的发展与进步。

2 结论与展望

综述机器学习方法在大气污染物浓度预测中的研究进展,系统评估传统机器学习方法、深度学习方法及其融合技术在该领域应用中的特征与性能。

1)传统机器学习方法是站点级短中期($t\leqslant72$ h)预测的实用技术。其中,SVR模型适配小样本、高维数据场景,能有效捕捉变量间非线性关联关系,但其计算效率较低且对异常值敏感;RF模型则抗过拟合能力突出,对数据缺失和异常值的鲁棒性

强,是站点级与区域级短中期预测的优选模型;而GBT系列模型在预测精度与计算效率之间实现了平衡,但对超参数调优敏感高度、可解释性较弱。这3类模型在主流研究中均表现出较高可靠性,可满足常规污染物预测的技术需求。

2)深度学习方法将是实现污染物浓度区域级、高分辨率预测的核心技术方向。其中,时间序列模型可精准捕捉污染物浓度的时序依赖特征,适用于短、中、长期等多周期预测;空间尺度模型聚焦污染物空间关联特征挖掘;时空融合模型实现了时间依赖性与空间关联性的同步建模,在市级、省域及跨区域污染物预测中表现出优异鲁棒性,尤其在极端污染事件预警中具有较高的命中率。

针对现有模型普遍存在的数据质量敏感、泛化能力有限及可解释性不足等问题,未来研究可从多源数据融合与模型优化升级双维度推进。一方面,整合地面监测数据、卫星遥感数据、气象再分析数据、排放源清单、地理空间数据及社会经济数据等多源信息,通过数据标准化、时空匹配和噪声消减等预处理技术,提升数据质量;同时结合迁移学习、联邦学习等方法,解决部分区域监测数据缺乏的问题,进而拓展模型的适用范围。另一方面,持续推进模型架构的轻量化、集成化与可解释性研究,通过单一模型优化、多模型融合手段,弥补传统机器学习与深度学习方法的性能短板,提高模型在跨区域、极端污染等复杂场景下的预测精度,为大气污染物精准防控提供更有力的技术支撑。

参考文献:

[1] Chikamoto Y, Zhang Wei, Hipps L, et al. Interannual variability and trends of summertime $PM_{2.5}$-based air quality in the Intermountain West [J/OL]. Environmental Research Letters, 2023, 18(4): 044032. https://doi.org/10.1088/1748-9326/acc6e0.

[2] Lu Yaling, Wang Yuan, Liao Yujie, et al. Public concern about haze and ozone in the era of their coordinated control in China [J/OL]. International Journal of Environmental Research and Public Health, 2023, 20(2): 911. https://doi.org/10.3390/ijerph20020911.

[3] Zhang Hongliang, Ying Qi. Source apportionment of airborne particulate matter in Southeast Texas using a source-oriented 3D air quality model [J]. Atmospheric Environment, 2010, 44(29): 3547-3557.

[4] Chen Fanglin, Chen Zhongfei. Cost of economic growth: Air pollution and health expenditure [J/OL]. Science of The Total Environment, 2021, 755: 142543. https://doi.org/10.1016/j.scitotenv.2020.142543.

[5] Rentschler J, Leonova N. Global air pollution exposure and poverty [J/OL]. Nature Communications, 2023, 14:4432. https://doi.org/10.1038/s41467-023-39797-4.

[6] Su I F, Chung Y C, Lee C, et al. Effective PM2.5 concentration forecasting based on multiple spatial-temporal GNN for areas without monitoring stations [J/OL]. Expert Systems with Applications, 2023, 234: 121031. https://doi.org/10.1016/j.eswa.2023.121074.

[7] Wen Wei, Shen Song, Liu Lei, et al. Comparative analysis of $PM_{2.5}$ and O_3 source in Beijing using a chemical transport model [J/OL]. Remote Sensing, 2021, 13(17): 3331. https://doi.org/10.3390/rs13173457.

[8] Žabkar R, Rakovec J, Koracin D. The roles of regional accumulation and advection of ozone during high ozone episodes in Slovenia: A WRF/Chem modelling study [J]. Atmospheric Environment, 2011, 45(5): 1192-1202.

[9] García Nieto P J, Sánchez Lasheras F, García Gonzalo E, et al. PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study [J]. Science of the Total Environment, 2018, 621: 753-761.

[10] Zhou Zheng, Qiu Cheng, Zhang Yufan. A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models [J/OL]. Scientific Reports, 2023, 13:22420. https://doi.org/10.1038/s41598-023-49899-0.

[11] Ghimire S, Deo R C, Jiang Ningbo, et al. Explainable deep learning hybrid modeling framework for total suspended particles concentrations prediction [J/OL]. Atmospheric Environment, 2025, 347: 121079. https://doi.org/10.1016/j.atmosenv.2025.121079.

[12] Li Chi, van Donkelaa A, Hamme M S, et al. Reversal of trends in global fine particulate matter air pollution [J/OL]. Nature Communications, 2023, 14: 5349. https://doi.org/10.1038/s41467-023-41086-z.

[13] Liao Qi, Zhu Mingming, Wu Lin, et al. Deep learning for air quality forecasts: A review [J]. Current Pollution Reports, 2020, 6(4): 399-409.

[14] Gu Haonan, Zhang Yinqiao, Shi Xuyuan, et al. Enhanced $PM_{2.5}$ prediction with a dynamic wind-driven graph fusion model incorporating inter-station pollutant transport [J/OL]. Atmospheric Environment, 2025,343:1210122. https://doi.org/10.1016/j.atmosenv.2024.121012.

[15] Gao Ziqi, Do K, Li Zongrun, et al. Predicting $PM_{2.5}$ levels and exceedance days using machine learning methods [J/OL]. Atmospheric Environment, 2024, 323: 120396. https://doi.org/10.1016/j.atmosenv.2024.120396.

[16] Peng Zezhi, Zhang Bin, Wang Diwei, et al. Application of machine learning in atmospheric pollution research: A state-of-art review [J/OL]. Science of The Total Environment, 2024, 910: 168588. https://doi.org/10.1016/j.scitotenv.2023.168588.

[17] Liu Zuhan, Huang Xin, Wang Xing. $PM_{2.5}$ prediction based on modified whale optimization algorithm and support vector regression [J/OL]. Scientific Reports, 2024, 14: 23296. https://doi.org/10.1038/s41598-024-74122-z.

[18] Sachdeva S, Kaur R, Kimmi, et al. Meteorological AQI and pollutants concentration-based AQI predictor [J]. International Journal of Environmental Science and Technology, 2023, 21(5): 4979-4996.

[19] Bahadure N B, Gonge S, Parashar D, et al. SVR machine learning and SARIMA-based air quality index classification and forecasting system [J/OL]. Discover Applied Sciences, 2025, 7: 994. https://doi.org/10.1007/s42452-025-07327-0.

[20] Wang Jujie, Xu Wenjie, Zhang Yue, et al. A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization [J/OL]. Chaos, Solitons & Fractals, 2022, 158: 12098. https://doi.org/10.1016/j.chaos.2022.112098.

[21] Houssein E H, Mohamed M, Younis E M G, et al. A hybrid harris hawks optimization with support vector regression for air quality forecasting [J/OL]. Scientific Reports, 2025, 15: 2275. https://doi.org/10.1038/s41598-025-86275-6.

[22] Castelli M, Clemente F M, Popovič A, et al. A machine learning approach to predict air quality in california [J]. Complexity, 2020, 332: 1-23.

[23] Pan Qilong, Harrou F, Sun Ying. A comparison of machine learning methods for ozone pollution prediction [J/OL]. Journal of Big Data, 2023, 10: 63. https://doi.org/10.1186/s40537-023-00748-x.

[24] 施光耀,杨思琪,张劲松,等.基于机器学习算法的高海拔地区臭氧影响因素重要性分析[J].宁夏大学学报(自然科学版),2024,45(2):196-202.

[25] Liu Huazhen, Liu Junfeng, Liu Ying, et al. Analysis of wintertime O_3 variability using a random forest model and high-frequency observations in Zhangjiakou: An area with background pollution level of the North China Plain [J/OL]. Environmental Pollution, 2020, 262: 114191. https://doi.org/10.1016/j.envpol.2020.114191.

[26] Wei Qing, Chen Yongqi, Zhang Huijin, et al. Simulation and prediction of $PM_{2.5}$ concentrations and analysis of driving factors using interpretable tree-based models in Shanghai, China [J/OL]. Environmental Research, 2025, 270: 121003. https://doi.org/10.1016/j.envres.2025.121003.

[27] Yun Dawei, Zheng Bing, Wu Haiwei, et al. Improving the short-term prediction of dissolved carbon monoxide using a combination of LightGBM and metaheuristic algorithms [J/OL]. Journal of Environmental Chemical Engineering, 2024, 12(5): 114043. https://doi.org/10.1016/j.jece.2024.114043.

[28] Zhong Junting, Zhang Xiaoye, Gui Ke, et al. Robust prediction of hourly $PM_{2.5}$ from meteorological data using LightGBM [J/OL]. National Science Review, 2021, 8(10): nwaa307. https://doi.org/10.1093/nsr/nwaa307.

[29] Ma Jinghui, Zhang Renhe, Xu Jianming, et al. MERRA-2 $PM_{2.5}$ mass concentration reconstruction in China's mainland based on LightGBM machine learning [J/OL]. Science of The Total Environment, 2022, 827: 154363. https://doi.org/10.1016/j.scitotenv.2022.154363.

[30] Li Jiantao, An Xingqin, Li Qingyong, et al. Application of XGBoost algorithm in the optimization of pollutant concentration [J/OL]. Atmospheric Research, 2022, 276: 106238. https://doi.org/10.1016/j.atmosres.2022.106238.

[31] Li Xiang, Peng Ling, Yao Xiaojing, et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation [J]. Environmental Pollution, 2017, 231: 997-1004.

[32] Xie Ziang, Chow C O, Chau J H, et al. Multi-pollutant air quality forecasting using bidirectional attention and multi-scale temporal networks [J/OL]. Frontiers in Environmental Science, 2025, 13: 1623630. https://doi.org/10.3389/fenvs.2025.1623630.

[33] Wu Zechuan, Tian Yuping, Li Mingze, et al. Prediction of air pollutant concentrations based on the long short-term memory neural network [J/OL]. Journal of Hazardous Materials, 2024, 465: 133099. https://doi.org/10.1016/j.jhazmat.2023.133099.

[34] Kurth T, Subramanian S, Harrington P, et al. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators [C/OL]// HUEBL A. Proceedings of the platform for advanced scientific computing conference, 2023. NY: Association for Computing Machinery.

https://doi.org/10.1145/3592979.3593412.

[35] Wu Jiaqi, Xu Lili, Fan Shurui, et al. A spatiotemporal graph attention network for $PM_{2.5}$ forecasting using multi-source data [J]. Air Quality, Atmosphere & Health, 2025, 18: 3037-3052.

[36] Li Xiang, Peng Ling, Hu Yuan, et al. Deep learning architecture for air quality predictions [J]. Environmental Science and Pollution Research, 2016, 23: 22408-22417.

[37] Zhan Yu, Luo Yuzhou, Deng Xunfei, et al. Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm [J]. Atmospheric Environment, 2017, 155: 129-139.

[38] Wang Shibao, mc Gibbon J, Zhang Yanxu. Predicting high-resolution air quality using machine learning: Integration of large eddy simulation and urban morphology data [J/OL]. Environmental Pollution, 2024, 344: 123371. https://doi.org/10.1016/j.envpol.2024.123371.

[39] Chen M H, Chen Y C, Chou T Y, et al. $PM_{2.5}$ concentration prediction model: A CNN-RF ensemble framework [J]. International Journal of Environmental Research and Public Health, 2023, 20(5): 4467. https://doi.org/10.3390/ijerph20054077.

[40] Bronstein M M, Bruna J, Lecun Y, et al. Geometric deep learning: Going beyond euclidean data [J]. IEEE Signal Processing Magazine, 2017, 34(4): 18-42.

[41] Qi Yanlin, Li Qi, Karimian H, et al. A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory [J]. Science of The Total Environment, 2019, 664: 1-10.

[42] Liao Haibin, Wu Mou, Yuan Li, et al. $PM_{2.5}$ prediction based on dynamic spatiotemporal graph neural network [J]. Applied Intelligence, 2024, 54: 11933-11948.

[43] Mandal S, Boppani S, Dasari V, et al. A bivariate simultaneous pollutant forecasting approach by unified spectro-spatial graph neural network (USSGNN) and its application in prediction of O- and NO_2 for New Delhi, India [J/OL]. Sustainable Cities and Society, 2024, 114: 105741. https://doi.org/10.1016/j.scs.2024.105741.

[44] Wen Congcong, Liu Shufu, Yao Xiaojing, et al. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction [J]. Science of The Total Environment, 2019, 654: 1091-1099.

[45] Yuan Jiang, Hua Dengxin, Wang Yufeng, et al. Attention mechanism based CNN-LSTM hybrid deep learning model for atmospheric ozone concentration prediction [J/OL]. Scientific Reports, 2025, 15: 21260. https://doi.org/10.1038/s41598-025-05877-2.

[46] Hu Feng, Xie Pinhua, Xu Jin, et al. Mapping regional meteorological processes to ozone variability in the North China Plain and the Yangtze River Delta, China [J]. Environmental Science & Technology, 2025, 59(19): 9632-9643.

[47] Bai Xuesong, Zhang Na, Cao Xiaoyi, et al. Prediction of $PM_{2.5}$ concentration based on a CNN-LSTM neural network algorithm [J/OL]. PeerJ, 2024, 12: e17811. https://doi.org/10.7717/peerj.17811.

[48] Wu Zonghan, Pan Shirui, Chen Fengwen, et al. A comprehensive survey on Graph Neural Networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[49] Zhao Qin, Liu Jiajun, Yang Xinwen, et al. Spatiotemporal $PM_{2.5}$ forecasting via dynamic geographical Graph Neural Network [J/OL]. Environmental Modelling & Software, 2025, 186: 106351. https://doi.org/10.1016/j.envsoft.2025.106351.

[50] Mandal S, Thakur M. A city-based $PM_{2.5}$ forecasting framework using Spatially Attentive Cluster-based Graph Neural Network model [J/OL]. Journal of Cleaner Production, 2023, 405: 137036. https://doi.org/10.1016/j.jclepro.2023.137036.

[51] Alawi O A, Kamar H M, Alsuwaiyan A, et al. Temporal trends and predictive modeling of air pollutants in Delhi: A comparative study of artificial intelligence models [J/OL]. Scientific Reports, 2024, 14: 30957. https://doi.org/10.1038/s41598-024-82117-z.

[52] Lim B, Zohren S. Time-series forecasting with deep learning: A survey [J/OL]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2021, 379 (2194): 20200209. https://doi.org/10.1098/rsta.2020.0209.

[53] Becerra-Rico J, Aceves-Fernández M A, Esquivel-Escalante K, et al. Airborne particle pollution predictive model using gated recurrent unit (GRU) deep neural networks [J]. Earth Science Informatics, 2020, 13(3): 821-834.

[54] Guyu Z, Yang Xiaoyuan, Shi Jiansen, et al. A $PM_{2.5}$ spatiotemporal prediction model based on mixed graph convolutional GRU and self-attention network [J/OL]. Environmental Pollution, 2025, 368: 125748. https://doi.org/10.1016/j.envpol.2025.125748.

[55] Wang Senzhang, Cao Jiannong, Yu P S. Deep learning for spatio-temporal data mining: A Survey [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(8): 3681-3700.

[56] Reichstein M, Camps-Valls G, Stevens B, et al. Deep learning and process understanding for data-driven earth system science [J]. Nature, 2019, 566 (7743): 195-204.

[57] Ma Jun, Cheng J C P, Lin Changqing, et al. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques [J/OL]. Atmospheric Environment, 2019, 214: 116885. https://doi. org/10.1016/j. atmosenv. 2019. 116885.

[58] Ye Yao, Cao Yong, Dong Yibo, et al. A graph neural network and transformer-based model for $PM_{2.5}$ prediction through spatiotemporal correlation [J/OL]. Environmental Modelling & Software, 2025, 191: 106501. https://doi. org/10.1016/j. envsoft. 2025. 106501.

Research Progress of Machine Learning in Predicting Atmospheric Pollutant Concentrations

ZHOU Shijie[1], *WANG Gang*[1*], *ZHAO Na*[2]

(1. College of Chemistry and Chemical Engineering, China University of Petroleum(East China), Qingdao 266580, China;
2. Environmental Research Institute, Shandong University, Qingdao 250100, China)

Abstract: Accurate prediction of atmospheric pollutant concentrations is essential for effective air pollution prevention and control. Traditional predictive methods possess inherent limitations in computational cost, nonlinear fitting capability, and generalization performance, which challenge their ability to meet the high-precision forecasting requirements in complex atmospheric environments. Recent advancements in machine learning technology offer promising new solutions in this arena. This paper systematically reviews machine learning methods employed for predicting atmospheric pollutant concentrations, classifying them into two primary categories: traditional machine learning methods (including support vector regression, random forests, and gradient boosting trees) and deep learning methods. Within the deep learning category, models are further classified based on task type into time series models, spatial scale models, and spatiotemporal fusion models. The study conducts a comprehensive comparative analysis of these models across several dimensions, including model mechanisms, predictive performance, applicable scenarios, and their respective advantages and disadvantages. The results indicate that traditional machine learning methods exhibit stable performance in short-to-medium-term ($t \leqslant$ 72 h) predictions at the station scale, demonstrating good applicability. Among deep learning methods, time series models can effectively improve prediction accuracy across multiple temporal scales, spatial scale models improve the capacity to capture spatial correlation features, and spatiotemporal fusion models facilitate the synergistic modeling of temporal dependencies and spatial correlations, showcasing superior overall predictive performance. Future developments in this field will likely focus on the integration of multi-source data, the creation of lightweight and integrated model architectures, and the generation of interpretable prediction results. Additionally, approaches such as transfer learning and federated learning may be employed to tackle the issue of data scarcity in certain regions, thereby expanding the applicability and predictive accuracy of models in scenarios involving cross-regional analyses and extreme pollution events. These advancements will provide robust technical support for the precise prevention and control of air pollution.

Key words: air pollutant prediction; traditional machine learning; deep learning

(责任编辑 高继红)